INCITE Introduction To Argonne’s BGL System

Susan Coghlan
High Performance Computing Manager
Mathematics and Computer Science Division
Argonne National Laboratory

3/02/06

Argonne National Laboratory is managed
The University of Chicago for the U.S. Department
Ener

Overview

- Configuration
- Preparing Codes
- Running Jobs
- Tools - Debugging and Others
- Help

Configuration Details

- Login servers [4]
  - Compile and submit jobs
  - bgl.mcs.anl.gov -> 2 servers DNS round-robin
  - login[1-4].bgl.mcs.anl.gov
- Service Node [1]
  - All jobs are started from the service node
  - It must be able to see the executable and starting directory (cwd)
  - Users have restricted shells on this server
- I/O nodes [32]
  - 1/32 IO node/compute node ratio
  - Computes are mapped to a specific IO node
  - ssh access allowed thru ZeptoOS kernel
- Compute nodes [1024]
  - No direct access
- Storage servers [20]
  - No user access

I/O on BGL

- Home directory
  - /bgl/home1/<username> (Aliased to /home/<username>)
  - Visible on: login servers, I/O nodes, computes, Service Node
  - Limited space (please watch usage)
  - Backed up nightly
  - Not a good idea to use for large quantity of accesses during job runs
- Local disk
  - /sandbox - **only** on the login servers, do not use for actual jobs
  - Scratch space - **not backed up!**
  - No local disk available on computes or I/O nodes
- Data
  - /pvfs/<username>
  - Visible on: login servers, I/O nodes, computes
  - Not visible to Service Node, so, no exec and no stderr/stdout files
  - **Not backed up!**

Building executables

- MPI wrappers (easiest): mpi<language>.<compiler>

mpicc.ibm	mpicxx.ibm	mpif77.ibm	mpif90.ibm
mpicc.gnu	mpicxx.gnu	mpif77.gnu	

  ex: mpicc.ibm -o HelloWorld.rts HelloWorld.c

- Be careful about mpicc vs mpicc.ibm
  - mpicc, mpicxx, mpif77 are IBM shipped wrappers that use the gnu compilers and have some problems
- Direct compiler and library linking also possible:
  - *Ex: /opt/ibmcmp/xlf/9.1/bin/blrts_xlf*
  - details in the **Hints & Tips** handout
- Optimizations
  - details in the **Hints & Tips** handout
  - advisors will assist with optimizations beyond the standard set

How our BGL configuration affects jobs

- 1 I/O node for each 32 compute nodes, hardwired, means minimum partition size of 32 nodes
- Partition sizes: 32, 64, 128, 256, 512, 1024 nodes
- Smaller partitions are enclosed inside of larger ones: *once a job is running on one of the smaller partitions, no jobs can run on the enclosing larger partitions*
- Not all partitions are available at all times
- Default BGL configuration: (1) 512 node partition, remaining split between default queue and short queue
- Processes are spread out in pre-defined mapping, sophisticated mappings possible with a map file
- Use ‘***bgl-listblocks***’ look at partitions, both active and non-active

Resource Mgr and Job Scheduler

- Cobalt - locally developed
- Standard commands, but prefaced with a ‘c’:
  - cqsub: submit jobs
  - cqstat: check job status
  - cqdel: delete jobs
- FIFO based, with some exceptions
- Queues
  - **default** - no need to specify
  - **short** - only jobs with <= 64 nodes and <= 30 minutes long
- Reservations
  - Required for anything larger than 512 nodes
  - Your production runs will need to be under reservations
  - Please contact support when you are ready to make production runs
  - Preventative maintenance reservation: Each Monday at 5pm

Submitting Jobs

- Examples for HelloWorld.rts executable:
  - *cqsub -q short -t 10 -n 32 HelloWorld.rts*
    - Will run in short queue
    - Will end after 10 minutes or once the executable exits whichever comes first
    - Will run on 32 nodes, 32 processors
    - Output will be stored in <jobid>.output and <jobid>.error

  Warning: don't specify -t less than 5 minutes
  - *cqsub -t 50 -n 128 -c 256 -m vn HelloWorld.rts*
    - Will run in default queue
    - Will run on 128 nodes, 256 processors
    - Warning: -m vn **required**
  - *cqsub -t 50 -n 256 -m vn HelloWorld.rts*
    - Will run on 256 nodes with 512 processors (due to -m vn)
- '*man cqsub'* for details about possible options

Why doesn't my job run?

- Possible causes:
  - Pending reservation
  - No partitions available
  - Wrong queue
  - Partitions not freed
- Use '*cqstat*' to see both running and waiting jobs
  - '*cqstat -f*' for more complete details (queue, etc)
  - Status: Q waiting, R running
- *showres*: show all defined reservations (pending and not yet deleted)
- *partlist*: show online partitions and status (sort of)
- Sometimes a job disappears from queue but is still holding a partition - '*bgl-listblocks*' can show if a partition is still allocated, '*bgl-listjobs*' will show jobs that BGL believes are still running

My job is no longer in the queue, but I don’t think it ran successfully…

- First place to look: STDERR file <jobid>.error
  - Sometimes the error messages are obscure - send mail to support
  - Note: Two job ids - Cobalt and BGL, both are important
- Are there any core files?
  - core.<node#>
  - ascii files, if you need help interpreting send mail to support
- Can you run a simple HelloWorld successfully?
  - If not, have you changed your dot files?
  - Are you forwarding X thru your ssh?
- Are your CWD and executable within your home directory space?
- Use print statements, but be aware that I/O is **very** buffered
- If all else fails, there is a limited version of gdb
  - You will need to request a partition for direct running of your job (I.e. not thru Cobalt)

Tools: Debugging & Others

- GDB - method of last resort, you will need to work with support
- Most tools are under: /soft/tools
- Heap/stack memory collision protection/tracking
- Tracing 'exit’ and 'abort’
- Libraries:
  - BLAS, LAPack
  - Mass, MassV
  - ESSL - very old version
  - FFTW
  - hdf5, netcdf
  - PETSc
  - Profiling
  - TAU
  - "-qdebug=function_trace”
- Your advisors will provide more details
- More on profiling tomorrow

Help!

- System issues (e.g. jobs not being scheduled, access problems, reservation requests,system not responding, etc.)

  email to support@bgl.mcs.anl.gov

- General BG/L questions and problems (e.g. what optimization flags work best, what libraries are available, how mapping works)

  email to discuss@bgl.mcs.anl.gov

- Resources:
  - BGL Hints & Tips document online:

    bgl.mcs.anl.gov:/software/common/doc/BGL-Hints-Tips.txt
  - BG/L wiki: http://wiki.bgl.mcs.anl.gov
  - BGL web pages: http://bgl.mcs.anl.gov

    Good starter page: http://bgl.mcs.anl.gov/Documentation/ (has links to IBM redbooks, new users guide, etc)